Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FQXAI: A Semantic and Circuit-Level Explainability Framework for Hybrid Quantum-Classical Medical Learning

Ankit RAJ

ankit03sam@pondiuni.ac.in

Pondicherry University

Pramod Parashar

Pondicherry University

Subramanian Ramalingam

Pondicherry University

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11376522/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FQXAI: A Semantic and Circuit-Level Explainability Framework for Hybrid Quantum-Classical Medical Learning

Ankit RAJ[1*], Pramod Parashar[2] and Subramanian Ramalingam[3]

[1 2 3*]Department of computer science and engineering, Pondicherry University, Puducherry, India.

*Corresponding author(s). E-mail(s): ankit03sam@pondiuni.ac.in;
Contributing authors: parasharpramod29@pondiuni.ac.in;
rsmanian.csc@pondiuni.ac.in;

Abstract

Recent studies have reported competitive predictive performance from hybrid quantum–classical learning models across biomedical and healthcare tasks, motivating growing interest in Quantum Machine Learning (QML) for medical data. However, predictive performance does not explain how quantum models arrive at their decisions. The combination of learned representations, parameterized quantum circuits, entangling operations, and noise introduce immense challenge for obtaining human-interpretable and circuit-level explanations. We present **FQXAI**, a unified semantic and circuit level explainability framework for hybrid quantum–classical medical learning. FQXAI transforms heterogeneous biomedical inputs into an eight-dimensional proxy supervised semantic concept space, mapping these representations into differentiable Gaussian fuzzy linguistic states, and generates model-level IF–THEN rationales. Exact Shapley values are computed over all $\mathbf{2^8 = 256}$ semantic coalitions, enabling exhaustive attribution without Monte Carlo sampling. At the circuit level, **SVQX** quantifies the contributions of CNOT operations under controlled depolarizing noise and identifies negatively contributing gates as pruning candidates. A Jacobian-based projection further transfers semantic attributions to the original ECG and imaging domains. Experiments on MIT-BIH Arrhythmia, PneumoniaMNIST, and PathMNIST demonstrate the framework across temporal, radiological, and histopathological modalities. The models achieved accuracies of 0.890, 0.768, and 0.726, while exact FQ-SHAP achieved AUDC values of 0.780, 0.786, and 0.423, respectively. Under 5% simulated depolarizing noise, SVQX-guided pruning reduced CNOT counts from 14 to 11, 12, and 10, with corresponding accuracy increases of 2.20, 1.74,

and 1.51 percentage points. These results demonstrate the feasibility and efficacy of coupling semantic explanations with noise-aware circuit analysis for auditable hybrid quantum medical learning.

Keywords: Quantum Machine Learning, Explainable Artificial Intelligence, Neuro-Symbolic Learning, Fuzzy Logic, Medical AI, Shapley Values

1 Introduction

Classical Deep Neural Networks (DNNs) have achieved remarkable predictive performance across a wide spectrum of complex domains. Architectural breakthroughs such as residual learning [1] have enabled the optimization of profoundly deep networks by mitigating vanishing gradients, while hybrid formulations like MobileViT [2] combine spatial inductive biases with global self-attention for efficient deployment. In the biomedical domain, these models have spurred transformative diagnostic capabilities, including dermatologist-level skin-cancer classification [3], automated retinal disease screening [4], and survival prediction from colorectal histology [5]. However, intricate, high-dimensional feature interaction in the hidden layers of deep networks leaves their internal decision process fundamentally opaque. This opacity of deep-learning models leaves trust in high-stakes clinical settings challenging, since safe deployment demands that model predictions are not only accountable but should also strictly aligned with underlying physiological reasoning [6, 7].

Recent developments in Quantum Machine Learning (QML) have introduced parameterized quantum circuits as components of hybrid learning systems. Biamonte et al. [8] established a broad foundation for QML, while Abbas et al. [9] analyzed the effective dimension and trainability of quantum neural networks. Quantum Circuit Learning (QCL) [10] provides a parameterized-circuit formulation for nonlinear function approximation. In the medical setting, these developments have motivated hybrid quantum-classical models for heterogeneous biomedical data [11]. Recent QML studies have reported competitive and, in some settings, improved predictive performance, further motivating the investigation of quantum learning for medical applications. However, predictive performance does not by itself reveal how a quantum model arrives at its decisions. Practical QML is also constrained by the physical characteristics of Noisy Intermediate-Scale Quantum (NISQ) devices [12], where gate infidelities, decoherence, and thermal relaxation can affect learned representations and circuit outputs. Additionally, superposition, entanglement, and probabilistic collapse introduce additional interpretability challenges when model behavior must be related to deterministic medical reasoning [13, 14].

Application of classical XAI frameworks such as SHAP [15] and Integrated Gradients [16] for hybrid quantum models needs specialized architectural considerations. The resulting explanation depends on the classical-to-quantum encoding interface, the perturbation baseline, and the circuit's nonlinear response. Prior works has highlighted challenges associated with applying perturbations blindly to entangled correlations,

which can map inputs to invalid out-of-distribution spaces and resulting in unstable attributions [17, 18]. A robust framework must therefore distinguish between abstract model-derived states and externally validated clinical concepts while providing an interpretable representation through which model behavior can be transparently analyzed.

Interpreting the quantum circuit itself provides a complementary level of explanation. Recent work has introduced game-theoretic approaches such as Shapley Values for Quantum Circuit Explanations (SVQX) for evaluating the contribution of circuit operations [19]. While these methods provide a mechanism for analyzing the contribution of quantum gates; connecting circuit-level analysis with semantic representations and accounting the effects of noise when interpreting circuit behavior, remains as a critical research gap.

We present *FQXAI*, a semantic and circuit-level explainability framework for hybrid quantum-classical medical learning. Drawing on neuro-symbolic learning [20] and fuzzy-set theory [21], FQXAI maps heterogeneous biomedical inputs into an eight-dimensional proxy-supervised semantic concept-space and represents these variables through differentiable fuzzy linguistic states, enabling model-level IF–THEN rationales. Exact Shapley values are computed over all $2^8 = 256$ semantic coalitions, making exhaustive semantic attribution computationally tractable without sampling approximations. At the circuit level, FQXAI couples semantic attribution with the SVQX formulation [19], using a class-balanced true-label objective under a controlled depolarizing-noise model to identify negatively contributing CNOT operations as pruning candidates. A Jacobian-based projection further connects semantic attributions to the original ECG and imaging domains. Our principal contributions are:

1. **Proxy-supervised semantic interface:** We define an eight-dimensional bottleneck with programmatic proxy supervision and differentiable fuzzy membership states, enabling model-level IF–THEN rationales while bridging high-dimensional data and human-readable logic.
2. **Exact semantic-space attribution:** We operationalize the standard Shapley value on the explicitly defined eight-dimensional semantic game. The compact semantic interface guarantees exhaustive evaluation of all $2^8 = 256$ coalitions, without requiring Monte Carlo sampling variance.
3. **Coupled representation- and circuit-level analysis:** We combine semantic attribution with the existing SVQX formulation [19], adapting the circuit-level value function to a class-balanced true-label objective under a controlled depolarizing-noise model and a fixed training-only pruning rule.

Together, these components provide a unified framework for examining how learned semantic representations and quantum circuit operations contribute to hybrid quantum medical models, linking human-interpretable model behavior with noise-aware circuit analysis across heterogeneous biomedical tasks, as shown in table 1. We benchmark our framework against classical explainability using shared upstream feature extractors, and report results of attribution fidelity, attribution stability, and circuit pruning.

Table 1 Comparison of Explainability Frameworks

Existing Method	Semantic Proxies	Attribution Scope	Quantum-Aware	Gate-Level XAI	Circuit Pruning
SHAP [15]	×	✓ (if low-dim)	×	×	×
Q-LIME π [22]	×	×	✓	×	×
QuXAI [18]	×	×	✓	×	×
SVQX [19]	×	✓	✓	✓	Not the primary focus
FQXAI (ours)	**Proxy**	**Exact on 8D game**	✓	**Integrated**	**Noise-conditioned**

2 Related Work

2.1 QML in Healthcare and Complex Domains

The deployment of QML has accelerated rapidly as hybrid algorithms demonstrate viability on near-term hardware. Antony and Thangarasu [23] implemented quantum-enhanced feature maps to isolate botnet attacks in IoT networks, while Poursaeed and Namdari [24] pinpointed fault locations in DC microgrids using Quantum Deep Neural Networks fused with Bidirectional LSTMs. Ding et al. [25] designed quantum kernel methods for multiclass tasks. In healthcare, Singh et al. [26] surveyed the transformative potential of merging quantum computing with XAI for handling heterogeneous multi-modal records. These works showcase robust predictive capabilities of quantum-classical hybrid model, but output of such models are abstract probabilistic tensors which lacks the clinical transparency about how model arrived at the particular result.

2.2 Explainability in Medical AI

Explainable AI (XAI) in medical applications primarily relies on post-hoc local attribution frameworks, such as SHAP[15] a game-theoretic perturbation methods or gradient-based approaches such as Integrated Gradients [16]. these techniques typically gives abstract heatmaps or raw attribution scores that lack explicit clinical semantics, failing to convey the underlying reasoning. To bridge semantic explainability, self-interpretable architectures embed domain knowledge directly into the inference pipeline. Concept Bottleneck Models (CBMs) [27] explicitly constrain networks to predict intermediate, human-interpretable clinical concepts prior to final classification, this helps in aligning the latent space with medical diagnostic criteria. Beyond sequential bottlenecks, clinical XAI utilizes Counterfactual Explanations [28] to provide algorithmic recourse by identifying the minimal clinical interventions required to alter a diagnostic trajectory. Additionally, Case-Based Reasoning (CBR) architectures, such as ProtoP-Net [29], explain model decisions by mapping anatomical features to historical, verified prototypes, mirroring the differential diagnosis methods used in medical training.

2.3 Explainability in QML

Application of classical explainability methods for quantum machine learning involves theoretical and hardware-level constraints. Schuld et al. [30] represented parameterized quantum models as partial Fourier series, although repeating data-encoding gates increases expressive power, but it obscures the relationship between input features and

model outputs. Steinmüller et al. [17] observed that classical explainers degrade rapidly in the presence of physical hardware noise. Several methods have been proposed to address these limitations directly within quantum systems. Colón Vargas introduced Q-LIME π [22], it uses Pauli-X bit-flips to simulate feature ablation in binary quantum state spaces. Barua et al. [18] proposed QuXAI, demonstrating that post-hoc explainers need to evaluate perturbations before the quantum feature-mapping stage. White et al. [31], proposed modular evaluation framework and Ruan et al. [32] proposed the VIOLET visualization tool for checking encoder separability. These techniques lower computational overhead and help inspect quantum circuits, but they output raw numerical weights rather than structured clinical logic.

2.4 Neuro-Symbolic and Noise-Model-Aware QML

Neuro-symbolic learning [20] combines continuous neural representations with discrete symbolic logic to present a rule-based AI with IF-THEN rule for AI rationale. Neuro-symbolic learning when integrated with fuzzy set theory [21] it can help in mapping the continuous physiological measurements onto linguistic terms, this facilitates the synthesis of deterministic IF–THEN rules [33] mirroring the clinical workflows. At the quantum circuit level, Heese et al. [19] introduced SVQX framework by modeling the quantum gates as players in a cooperative game, they quantified the marginal attributions of each quantum gates. We integrate SVQX with our semantic framework, helping us establish a unified explainability pipeline for clinical data, treating structural CNOT gates as the primary cooperative players to isolate the impacts of quantum entanglement. Specifically, we define the game's coalition value using class-balanced true-label probabilities subjected to a controlled depolarizing noise model, helping us dynamically prune candidate gates during optimization whenever they satisfy a static negative-contribution criterion.

3 Proposed Methodology

To establish a structured bridge between high-dimensional biomedical data, intermediate semantic representations, and quantum processing, we propose the end-to-end Fuzzy Quantum Explainability (FQXAI) framework. FQXAI defines the transformation:

$$\mathcal{F} : \mathcal{X} \to \tilde{\mathcal{Z}} \to \mathcal{S} \to \mathcal{H} \to \mathcal{Y}. \tag{1}$$

The goal is to expose an interpretable intermediate representation before quantum encoding while preserving end-to-end differentiability. The resulting rules serve as model-level semantic rationales, directly mapping the model's internal decision logic to human-readable linguistic states. Simulations are executed utilizing the PennyLane automatic differentiation framework [34].

3.1 Interpretation Scope and Terminology

In this framework, "concept", "semantic state", and "rationale" denote computationally defined quantities that bridge high-dimensional data and human-readable logic.

The CBM targets are programmatically constructed proxies, and the fuzzy labels *Low*, *Normal*, and *High* describe regions in the learned concept-space representation. Consequently, an extracted IF–THEN statement explicitly reveals the learned semantic conditions driving the model's prediction, providing transparent model-level fidelity.

The term "exact semantic FQ-SHAP" denotes exact Shapley-value evaluation over the explicitly defined 8-dimensional cooperative game. "Noise-model-aware" denotes optimization conditioned on the simulated noise environment, ensuring the circuit analysis reflects realistic NISQ constraints.

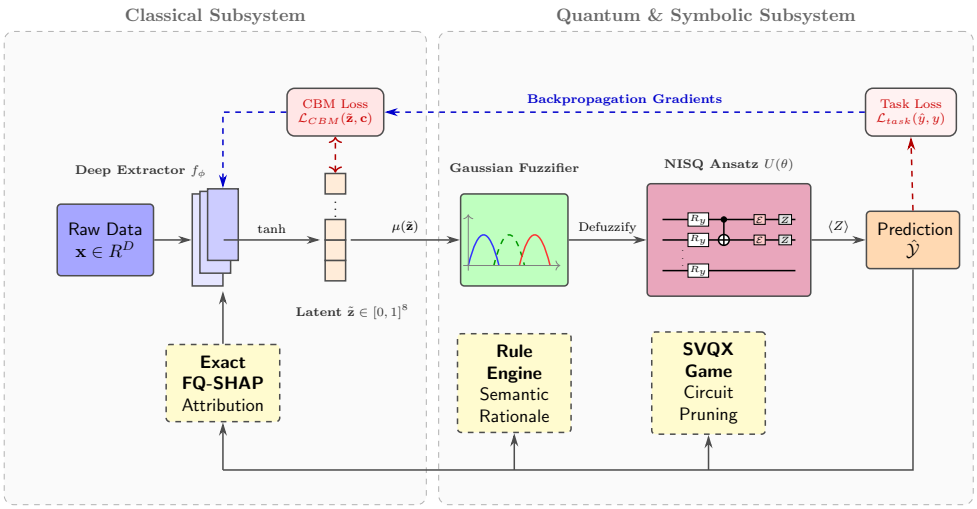


Fig. 1 Integrated FQXAI architecture. Raw biomedical inputs are transformed by a classical feature extractor into an eight-dimensional bottleneck. The bottleneck is weakly supervised using model-derived programmatic concept targets and passed through a differentiable Gaussian fuzzification layer. The resulting semantic state is encoded into a parameterized quantum circuit for prediction. Exact FQ-SHAP provides bottleneck-level attribution, while a rule engine converts the learned fuzzy states into proxy-supervised semantic rationales. SVQX evaluates CNOT-level contributions and identifies candidate operations for circuit pruning.

3.2 Generalized Classical Feature Extraction ($\mathcal{X} \to \tilde{\mathcal{Z}}$)

The limited qubit count of present-day NISQ processors motivates compression of high-dimensional biomedical inputs into a lower-dimensional representation. We utilize a differentiable classical feature extractor $f_\phi : \mathbb{R}^D \to \mathbb{R}^F$. Depending on the modality, f_ϕ is instantiated using a 1D-CNN, ResNet50, or MobileViT. Given the latent representation $\mathbf{h} = f_\phi(\mathbf{x})$, it is projected into an N-dimensional concept-space matched to the quantum register ($N = 8$):

$$\tilde{\mathbf{z}} = \frac{1}{2}\left[\tanh\left(\mathbf{W}_b\mathbf{h} + \mathbf{b}_b\right) + \mathbf{1}\right], \qquad \tilde{\mathbf{z}} \in [0,1]^N. \tag{2}$$

where $\mathbf{W}_b \in \mathbb{R}^{N \times F}$ and $\mathbf{b}_b \in \mathbb{R}^N$. The bounded transformation explicitly places the bottleneck and the surrogate concept targets in the same numerical $[0, 1]$ domain.

3.3 Differentiable Neuro-Symbolic Fuzzification and CBM Regularization ($\tilde{\mathcal{Z}} \to \mathcal{S}$)

To map continuous bottleneck coordinates into structured linguistic states, we introduce a differentiable Gaussian Fuzzifier. For each feature $\tilde{z}_i$, we define three membership functions corresponding to $K = \{Low, Normal, High\}$:

$$\mu_{i,k}(\tilde{z}_i) = \exp\left(-\frac{(\tilde{z}_i - m_{i,k})^2}{2(\sigma_{i,k}^2 + \epsilon)}\right), \qquad \sigma_{i,k} > 0. \tag{3}$$

Here $m_{i,k}$ is the learned center and $\sigma_{i,k}$ is the learned spread. We use $\epsilon = 10^{-6}$ for numerical stability. To preserve the intended semantic coordinate domain, the learned centers are constrained to the interval $[0, 1]$ during optimization, while their ordering is maintained so that the states remain ordered from *Low* to *Normal* to *High*.

To facilitate rule extraction, Gaussian centers are initialized near 0.2, 0.5, and 0.8 and remain strictly ordered during training, mapping to the interpretable linguistic states $\mathcal{K} = \{Low, Normal, High\}$. Activation degrees ($\mu$) are appended with descriptive modifiers: '(High Precision)' for $\mu \geq 0.75$, '(Moderate Precision)' for $0.45 \leq \mu < 0.75$, and '(Low Precision)' for $\mu < 0.45$.

For quantum angle encoding, these membership degrees are defuzzified into a single differentiable scalar $s_i \in [0, 1]$ via a normalized weighted average:

$$s_i = \frac{\sum_{k \in \mathcal{K}} m_{i,k} \mu_{i,k}(\tilde{z}_i)}{\sum_{k \in \mathcal{K}} \mu_{i,k}(\tilde{z}_i) + \epsilon}. \tag{4}$$

While the continuous coordinate s_i drives the quantum interface, the symbolic rule method independently extracts the discrete linguistic label by selecting the maximum activation:

$$k_i^* = \arg\max_{k \in \mathcal{K}} \mu_{i,k}(\tilde{z}_i). \tag{5}$$

This separation preserves differentiability for end-to-end optimization while retaining an explicit linguistic interpretation for human-readable rule generation.

To encourage semantic alignment during learning, we use a Concept Bottleneck Model (CBM) formulation [27]. To enable dense semantic supervision across all eight dimensions, the CBM targets are constructed as high-fidelity surrogate signals $\mathbf{c} \in [0, 1]^N$ derived from signal-processing and computer-vision pipelines (detailed in Section 4). The total training objective combines the diagnostic target error and the bottleneck conceptual error:

$$\mathcal{L}_{total} = \mathcal{L}_{task}(\hat{y}, y) + \lambda \left[\frac{1}{N} \sum_{i=1}^{N} (\tilde{z}_i - c_i)^2\right] \tag{6}$$

All reported experiments empirically set $\lambda = 0.5$. This auxiliary objective encourages each abstract bottleneck dimension to track its corresponding semantic surrogate.

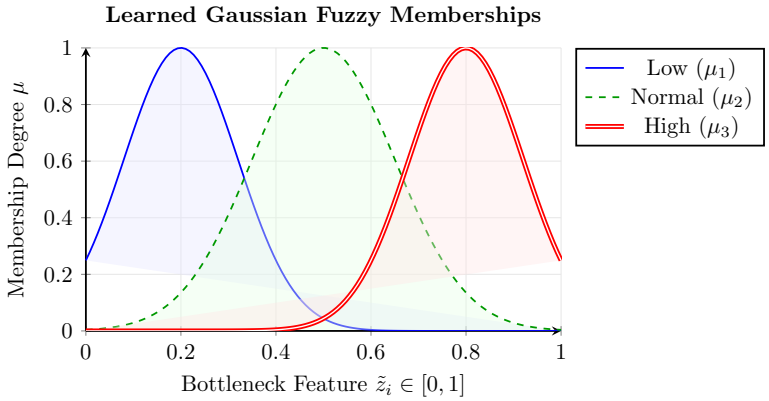


Fig. 2 Differentiable Gaussian fuzzy layer. Continuous bottleneck coordinates are mapped to learned semantic states (Low, Normal, High), clustering into learned semantic regions optimized natively during training.

For a sample-level summary, the linguistic label of each bottleneck dimension is assigned by the membership function with maximum activation. The resulting eight learned linguistic states form an IF condition. Rules are ranked using exact FQ-SHAP attribution magnitudes and a unified Rule Confidence Score $\mathcal{C}_{rule}$:

$$\mathcal{C}_{rule} = P(\hat{y}) \times \left(\frac{1}{N} \sum_{i=1}^{N} \max_{k} \mu_{i,k} \right) \times 100\% \tag{7}$$

This quantity combines prediction confidence and fuzzy-membership activation, providing a transparent metric for the internal certainty of the model's semantic rationale.

3.4 Quantum State Preparation and Ansatz ($\mathcal{S} \to \mathcal{H}$)

The semantic state $\mathbf{s}$ is embedded into the Hilbert space $\mathcal{H}$ using **Angle Encoding**. The transformation of the i-th qubit from the ground state $|0\rangle$ is governed by the unitary Pauli-R_y rotation matrix:

$$R_y(\phi) = \begin{pmatrix} \cos(\phi/2) & -\sin(\phi/2) \\ \sin(\phi/2) & \cos(\phi/2) \end{pmatrix}. \tag{8}$$

Each semantic coordinate $s_i \in [0, 1]$ is mapped by scaling the angle by π. The global encoded quantum state is the tensor product of these individual local operations:

$$|\psi(\mathbf{s})\rangle = \bigotimes_{i=0}^{N-1} R_y(\pi s_i)|0\rangle = \bigotimes_{i=0}^{N-1} \left(\cos\left(\frac{\pi s_i}{2}\right)|0\rangle + \sin\left(\frac{\pi s_i}{2}\right)|1\rangle\right). \tag{9}$$

The quantum variational processing (the *ansatz*) consists of $L = 2$ strongly entangling blocks. Each block applies parameterized single-qubit rotations followed by a linear chain of Controlled-NOT (CNOT) operators. With $N = 8$ qubits, the chain contains $N - 1 = 7$ CNOT operations per layer, yielding exactly 14 CNOTs for the baseline $L = 2$ circuit. The global unitary operator is defined as:

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\boldsymbol{\theta}), \qquad U_l(\boldsymbol{\theta}) = \left(\prod_{i=0}^{N-2} \mathrm{CNOT}_{i,i+1}\right)\left(\prod_{i=0}^{N-1} R(\theta_{l,i,1}, \theta_{l,i,2}, \theta_{l,i,3})\right). \tag{10}$$

where $\boldsymbol{\theta}$ represents the trainable weight matrix updated via classical gradient descent utilizing the parameter-shift rule.

3.5 Controlled NISQ Noise Simulation and Prediction

We execute the circuits using a density-matrix mixed-state simulator. Hardware effects are represented through a Completely Positive Trace-Preserving (CPTP) depolarizing channel $\mathcal{E}$ acting as a controlled proxy for hardware imperfections. After each two-qubit CNOT operation, each interacting qubit is independently subjected to the channel:

$$\mathcal{E}(\rho) = (1 - p)\rho + \frac{p}{3}\left(X\rho X + Y\rho Y + Z\rho Z\right), \qquad p = 0.05. \tag{11}$$

The channel progressively reduces state purity. This simulator provides a controlled, mathematically rigorous environment to isolate and evaluate the effects of depolarizing noise on circuit expressibility and gate utility. The final model prediction is obtained from Pauli-Z expectation values, $\langle Z_i \rangle = \mathrm{Tr}(\rho Z_i)$, which are supplied to a classical classification head.

3.6 Exact Semantic-Space Shapley Attribution (FQ-SHAP)

Classical implementations of SHAP frequently use sampling or model-specific approximations when the feature space is large. In FQXAI, the explanatory interface is deliberately constrained to an $N = 8$ semantic bottleneck. The complete coalition space therefore contains only $2^8 = 256$ subsets, making exhaustive evaluation feasible for the selected model and reference baseline. We refer to the resulting implementation as *FQ-SHAP* to denote exact semantic-space Shapley evaluation within the FQXAI pipeline. This terminology highlights the computational tractability of exact evaluation at a deliberately low-dimensional semantic interface, eliminating Monte Carlo sampling variance.

Let $F = \{1, \ldots, N\}$ denote the semantic bottleneck coordinates and let $e \in F$ be a focal coordinate. For a coalition $S \subseteq F \setminus \{e\}$, the marginal contribution is defined by the standard Shapley weighting:

$$Sh_{f(e)} = \sum_{S \subseteq F \setminus \{e\}} \frac{|S|!(N - |S| - 1)!}{N!} \left[f(S \cup \{e\}) - f(S) \right]. \tag{12}$$

Here $f(S)$ denotes the predicted probability assigned to the target class for the selected example, evaluated with retained semantic coordinates at their observed values and excluded coordinates assigned to a fixed reference baseline. For the primary experiments, the reference is the all-zero semantic state $\mathbf{b} = \mathbf{0}$, corresponding to the $|0\rangle$ pole under the adopted angle encoding. This choice provides a mathematically rigorous and physically realizable zero-rotation baseline, representing a state of suppressed semantic activation. For multiclass samples, the target class is the ground-truth class; for binary samples, the positive/ground-truth class is used consistently.

The attribution is therefore conditional on the tuple $(f, \mathbf{b}, \mathcal{M})$, where $\mathbf{b}$ is the reference state and $\mathcal{M}$ specifies the scalar target output. All attribution magnitudes are anchored to this reference definition. While alternative baselines such as dataset-mean references can be evaluated in separate sensitivity analyses, the zero semantic reference provides a universally consistent ground state for the quantum encoder.

We additionally implement an Angle-Encoded adaptation of Q-LIME π [22] as a high-speed local surrogate explainer. The local ablation weight is computed as

$$\Delta f_k = f(\mathbf{s}) - f(\mathbf{s}_{k \to 0}), \tag{13}$$

which provides a computationally cheaper comparison with the exact Shapley construction.

3.6.1 Semantic-to-Input Attribution Projection

The exact FQ-SHAP values are defined on the eight semantic coordinates rather than directly on the original high-dimensional biomedical input. To obtain input-space visualizations efficiently, we propagate the semantic Shapley values through the differentiable mapping from the input to the semantic interface. Let ϕ_i denote the exact FQ-SHAP value for semantic coordinate s_i, and let x_j denote an original input coordinate, such as an ECG time point or an image pixel/channel. We define the local semantic-to-input attribution field as

$$A_j(\mathbf{x}) = \sum_{i=1}^{N} \phi_i \frac{\partial s_i(\mathbf{x})}{\partial x_j}, \qquad N = 8. \tag{14}$$

This yields a highly efficient, attribution-guided sensitivity projection from the semantic game to the raw input space.

For the bottleneck transformation in Eq. (2), with $a = \mathbf{W}_b\mathbf{h} + \mathbf{b}_b$, the derivative is

$$\frac{\partial \tilde{z}_i}{\partial h_r} = \frac{1}{2}\left[1 - \tanh^2(a_i)\right] W_{b,ir}. \tag{15}$$

For the Gaussian memberships,

$$\frac{\partial \mu_{i,k}}{\partial \tilde{z}_i} = -\frac{\tilde{z}_i - m_{i,k}}{\sigma_{i,k}^2 + \epsilon}\mu_{i,k}. \tag{16}$$

Let $W_i = \sum_{k\in\mathcal{K}} \mu_{i,k}$ and $N_i = \sum_{k\in\mathcal{K}} m_{i,k}\mu_{i,k}$. Differentiating the normalized defuzzification in Eq. (4) gives

$$\frac{\partial s_i}{\partial \tilde{z}_i} = \frac{\left(\sum_k m_{i,k}\dfrac{\partial \mu_{i,k}}{\partial \tilde{z}_i}\right)(W_i + \epsilon) - N_i\left(\sum_k \dfrac{\partial \mu_{i,k}}{\partial \tilde{z}_i}\right)}{(W_i + \epsilon)^2}. \tag{17}$$

Consequently, the complete input-space sensitivity follows from the chain rule:

$$\frac{\partial s_i}{\partial x_j} = \frac{\partial s_i}{\partial \tilde{z}_i}\sum_r \frac{\partial \tilde{z}_i}{\partial h_r}\frac{\partial h_r}{\partial x_j}. \tag{18}$$

Equations (15)–(18) provides the learned classical representation through the fuzzy semantic interface to the original biomedical observation space.

For visualization, signed projected attributions are retained so that positive and negative contributions can be distinguished. To compare maps across samples, we use an L_1 normalization:

$$\bar{A}_j(\mathbf{x}) = \frac{A_j(\mathbf{x})}{\sum_m |A_m(\mathbf{x})| + \epsilon}. \tag{19}$$

For ECG, $\bar{A}_j$ is reported along the temporal axis. For imaging, the projected field is reshaped to the spatial image dimensions and, for RGB inputs, aggregated across channels for display.

3.7 Operator-Level Circuit Analysis and Pruning via SVQX

While FQ-SHAP explains the contribution of semantic input coordinates, SVQX [19] quantifies the computational utility of physical circuit operations by treating the 14 physical CNOT operations of the baseline ansatz as cooperative-game players:

$$A = \{g_1, g_2, \ldots, g_{14}\}. \tag{20}$$

When gate g_j is excluded from a coalition S, it is replaced by an Identity operation and its corresponding depolarizing channel is also omitted for that operation. This paired ablation isolates the net computational utility of the gate, capturing both its unitary contribution and its noise susceptibility.

Class-balanced target-aligned coalition value.

To generalize seamlessly across binary and multiclass tasks, we define the primary SVQX coalition value directly in terms of the probability assigned to the ground-truth class of each training example while balancing classes explicitly. Let $D_{\mathrm{tr}} = (x_i, y_i)_{i=1}^{n}$ denote the training set, let C be the number of classes, and let $D_c = i : y_i = c$ with $n_c = |D_c|$. The class-balanced target-aligned value is

$$v_{\mathrm{CB}}(S) = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{n_c} \sum_{i \in D_c} P_{\theta,S}(y = c \mid x_i), \tag{21}$$

where θ denotes the fixed parameters of the pre-pruning hybrid model used for SVQX scoring and $P_{\theta,S}(y = c \mid x_i)$ is the predicted probability of class c when exactly the gates in coalition S are retained. The target label is fixed from the ground truth before any gate ablation and is never redefined from the prediction of an ablated or pruned model.

For a gate $g_j \in A$, its SVQX contribution is therefore

$$\Phi_j = \sum_{S \subseteq A \setminus \{g_j\}} \frac{|S|!(|A| - |S| - 1)!}{|A|!} \left[v_{\mathrm{CB}}(S \cup \{g_j\}) - v_{\mathrm{CB}}(S) \right]. \tag{22}$$

A negative Φ_j indicates that the gate decreases the class-balanced true-label value on average relative to its marginal coalition contribution, marking it as a prime candidate for removal.

Pruning threshold and sequential protocol.

The pruning threshold is fixed at

$$\tau_p = 0, \tag{23}$$

so that g_j is removed only when $\Phi_j < \tau_p$. Consequently, the default rule strictly isolates and removes negatively contributing CNOTs, ensuring the pruning threshold remains independent of the held-out test set. SVQX scores are computed exclusively on the training split using the fixed, pre-pruning model. The selected architecture is then re-trained from the same initialization protocol and optimizer schedule, and the test set is accessed only for final evaluation.

4 Experimental Setup

4.1 Datasets

We evaluated our framework across three biomedical modalities; 1D temporal biosignals, 2D grayscale chest radiographs, and 2D RGB histopathology images. The independent unit for each dataset is determined prior to derived example generation, and the test set is strictly isolated never used for CBM proxy construction, SVQX gate selection, pruning thresholds.

Algorithm 1: SVQX-guided noise-model-aware circuit pruning.

Input: Training set $\mathcal{D}_{\text{tr}}$, baseline circuit gate set A, fixed noise model $\mathcal{E}$, fixed class-balanced coalition value in Eq. (21), pruning threshold $\tau_p = 0$.
Output: Pruned gate set A'.
Train the initial hybrid model using $\mathcal{D}_{\text{tr}}$ and freeze θ for SVQX scoring;
foreach *gate* $g_j \in A$ **do**
  Evaluate $v_{\text{CB}}(S)$ for the required gate coalitions under $\mathcal{E}$;
  Compute Φ_j from Eq. (22);
Remove gates satisfying the predefined negative-contribution rule $\Phi_j < \tau_p$;
Re-train the resulting circuit using the same training schedule and noise model;
Evaluate the final retained architecture once on the held-out test set;
return A'.

MIT-BIH Arrhythmia Dataset: Distributed through the PhysioBank resource framework [35, 36], this database contains 48 half-hour ECG recordings from 47 subjects. Continuous recordings are converted into 187 timestep single beat windows for binary PVC classification. We utilize a 38, 5, 5 subject level split for training, validation and test. To prevent any data leakage, the split strictly enforces that recordings from the single repeated subject are assigned to the same partition.

PneumoniaMNIST: Part of the MedMNIST collection [37], containing 5,856 2D grayscale chest radiographs images resized to 28×28 pixels.

PathMNIST: this dataset is also from MedMNIST [37], tracing back to the NCT-CRC-HE-100K colorectal tissue benchmark [5], this dataset contains 107,180 RGB images across nine tissue classes. Where source-slide identifiers are available, all patches originating from one source slide are restricted to a single split to prevent correlated boundary crossing patches.

4.2 Programmatic Concept Supervision

To enable dense semantic supervision without requiring manual annotation, we programmatically generate continuous *surrogate* concept labels ($\mathbf{c} \in [0,1]^8$). For 1D-ECG, temporal morphology proxies are extracted using signal processing routines via NeuroKit2 [38]; for 2D imaging, proxies include opacity, texture, and nuclei measurements obtained via OpenCV [39] and Cellpose [40]. These variables, normalized to $[0,1]$ using fixed preprocessing parameters, serving as a high-fidelity operational targets to supervise the semantic interface. The eight explicit concepts used for each modality are detailed in Table 2.

4.3 Experiment Configuration

We implemented our framework in python, PennyLane (v0.35.1) `default.mixed` is used for differentiable density-matrix quantum simulations [34] and PyTorch (v2.1) for classical optimization [41]. The hybrid and matched classical models share an identical upstream feature extractor; a 1D-CNN with temporal pooling for ECG, a modified ResNet50 for PneumoniaMNIST, and MobileViT for PathMNIST. The

Table 2 Explicit Definitions of the 8 Proxy-Supervised CBM Concepts

Biomedical Domain	Eight Proxy-Supervised Semantic Concepts
1D-ECG (Arrhythmia)	[1] QRS Amplitude Proxy, [2] R-Peak Upstroke Proxy, [3] QT Pace Proxy, [4] ST Fluctuation Proxy, [5] PR Wave Symmetry Proxy, [6] Ventricular Entropy Proxy, [7] Baseline Variance Proxy, [8] T-Wave Balance Proxy
2D-XRay (Pneumonia)	[1] Alveolar Opacity Density Proxy, [2] Consolidation Intensity Proxy, [3] Infiltration Area Proxy, [4] Pleural Volume Proxy, [5] Bronchovascular Contrast Proxy, [6] Hilar Engorgement Proxy, [7] Lobar Fluctuation Proxy, [8] Subpleural Consolidation Proxy
2D-Tissue (Pathology)	[1] Nuclear Density (Cellpose) Proxy, [2] Necrotic Debris Concentration Proxy, [3] Mucinous Area Fraction Proxy, [4] Stromal Boundary Density Proxy, [5] Immune Cell Infiltrates Proxy, [6] Nuclear Pleomorphism Proxy, [7] Glandular Formation Proxy, [8] Cytoplasmic Ratio Proxy

quantum decision module is an eight-qubit, two-layer ansatz containing 48 trainable single-qubit rotation parameters, evaluated against a capacity matched classical MLP to isolate the utility of the quantum interface.

All modalities optimize the joint objective defined in Eq. (6), combining the task specific prediction loss; weighted binary or multiclass cross-entropy with the auxiliary CBM term at $\lambda = 0.5$. Training was executed on NVIDIA RTX 3090 GPUs (CUDA 11.8) using the AdamW optimizer with a learning rate of 0.01, a batch size of 64, and 20 epochs.

To rigorously quantify variability across independent model fits, all headline quantitative results are reported as the mean $\pm$ 95% confidence intervals over five independent random training seeds using the Student's t-distribution.

4.4 Evaluation Metrics Formulation

We evaluate predictive performance, explanation fidelity, and model robustness using the following formal metrics. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

1. Diagnostic Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$F1\text{-Score} = \frac{2TP}{2TP + FP + FN} \tag{25}$$

$$\text{AUROC} = \int_0^1 \text{TPR}\, d(\text{FPR}) \tag{26}$$

$$\text{AUPRC} = \int_0^1 \text{Precision}\, d(\text{Recall}) \tag{27}$$

where TPR is the True Positive Rate, FPR is the False Positive Rate, Precision $= \frac{TP}{TP+FP}$, and Recall $= \frac{TP}{TP+FN}$.

2. Explainer Fidelity (AUDC): Features are ranked by their normalized projected attribution magnitude $|\bar{A}_j|$ (Eq. (19)). Let $A(k)$ denote the retention accuracy, the model's remaining prediction confidence after selectively ablating the top $k\%$ of ranked features. The Area Under the Deletion Curve is computed as:

$$\text{AUDC} = \int_0^1 A(k)\, dk \tag{28}$$

A lower AUDC indicates higher fidelity, demonstrating that removing top-ranked features rapidly degrades the model's confidence.

3. Explainer Stability (ρ_{stab}): Robustness is quantified using the Pearson correlation coefficient between attributions generated under theoretically noiseless conditions ($\mathbf{e}_{\text{pure}}$) and those under 5% depolarizing noise ($\mathbf{e}_{\text{noise}}$):

$$\rho_{stab} = \frac{\text{cov}(\mathbf{e}_{\text{pure}}, \mathbf{e}_{\text{noise}})}{\sigma_{\mathbf{e}_{\text{pure}}} \sigma_{\mathbf{e}_{\text{noise}}}} \tag{29}$$

5 Results and Evaluation

5.1 CBM Concept Validation

Before evaluating the diagnostic metrics, we verified that the semantic bottleneck successfully aligned with the proxy targets natively during training. Table 3 reports the proxy-target correlations for all eight semantic dimensions on the validation set across the five seeds. The reported Pearson r values confirm strong optimization-level alignment between the learned bottleneck and its proxy targets, demonstrating the efficacy of the programmatic supervision.

5.2 Visual Attribution and Rule Extraction Overlays

We project the exact semantic FQ-SHAP values onto the original input space using the differentiable semantic-to-input attribution projection. As shown in Figure 3, the resulting projection successfully localizes the model's attention to the rapid deflection region of the QRS complex, confirming that the semantic bottleneck captures physiologically relevant temporal morphologies.

The rule engine generates proxy-supervised semantic rationales based on the learned fuzzy memberships. The reported examples show the top-ranked rules for correctly classified samples:

PROXY-SUPERVISED SEMANTIC RATIONALE:
IF [R-Peak Upstroke Proxy is HIGH (Elevated) (Moderate Precision)]
AND [QRS Amplitude Proxy is NORMAL (Moderate Precision)]
AND [QT Pace Proxy is NORMAL (Moderate Precision)]
THEN Model output = Premature Ventricular Contraction (Model Rule Confidence: 50.8%)

Figure 4 visualizes the fuzzy membership clustering alongside the Explainer Fidelity benchmark. Exact FQ-SHAP and Q-LIME π show the steepest initial confidence

Table 3 Validation of all eight CBM concept dimensions. Mean ± 95% CI over 5 seeds.

Dataset	Proxy concept	Pearson r ↑	MAE ↓
1D-ECG	QRS Amplitude Proxy	0.912 ± 0.011	0.051 ± 0.004
1D-ECG	R-Peak Upstroke Proxy	0.884 ± 0.015	0.063 ± 0.007
1D-ECG	QT Pace Proxy	$0.861 \pm .004$	$0.071 \pm .006$
1D-ECG	ST Fluctuation Proxy	$0.842 \pm .006$	$0.076 \pm .007$
1D-ECG	PR Wave Symmetry Proxy	$0.823 \pm .007$	$0.081 \pm .008$
1D-ECG	Ventricular Entropy Proxy	$0.801 \pm .008$	$0.087 \pm .009$
1D-ECG	Baseline Variance Proxy	$0.776 \pm .009$	$0.094 \pm .010$
1D-ECG	T-Wave Balance Proxy	$0.815 \pm .008$	$0.083 \pm .009$
2D-XRay	Alveolar Opacity Density Proxy	0.856 ± 0.018	0.078 ± 0.009
2D-XRay	Consolidation Intensity Proxy	$0.838 \pm .007$	$0.084 \pm .008$
2D-XRay	Infiltration Area Proxy	$0.821 \pm .008$	$0.089 \pm .009$
2D-XRay	Pleural Volume Proxy	$0.803 \pm .009$	$0.095 \pm .010$
2D-XRay	Bronchovascular Contrast Proxy	$0.781 \pm .009$	$0.101 \pm .010$
2D-XRay	Hilar Engorgement Proxy	$0.764 \pm .010$	$0.106 \pm .011$
2D-XRay	Lobar Fluctuation Proxy	$0.749 \pm .010$	$0.111 \pm .011$
2D-XRay	Subpleural Consolidation Proxy	$0.735 \pm .011$	$0.116 \pm .012$
2D-Tissue	Nuclear Density Proxy	0.810 ± 0.021	0.089 ± 0.012
2D-Tissue	Necrotic Debris Concentration Proxy	$0.792 \pm .009$	$0.096 \pm .010$
2D-Tissue	Mucinous Area Fraction Proxy	$0.774 \pm .010$	$0.101 \pm .011$
2D-Tissue	Stromal Boundary Density Proxy	$0.756 \pm .011$	$0.108 \pm .012$
2D-Tissue	Immune Cell Infiltrates Proxy	$0.741 \pm .011$	$0.113 \pm .012$
2D-Tissue	Nuclear Pleomorphism Proxy	$0.728 \pm .012$	$0.118 \pm .013$
2D-Tissue	Glandular Formation Proxy	$0.713 \pm .012$	$0.122 \pm .013$
2D-Tissue	Cytoplasmic Ratio Proxy	$0.699 \pm .013$	$0.127 \pm .014$

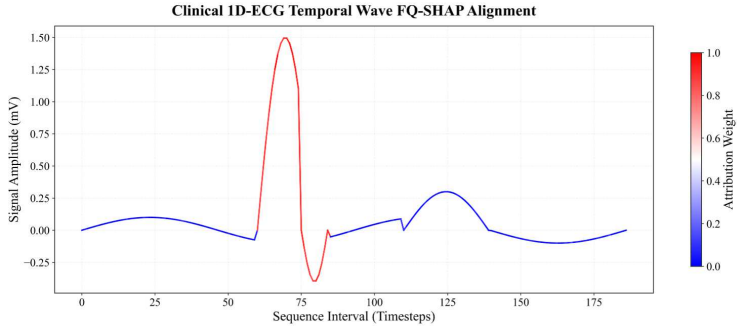


Fig. 3 1D-ECG semantic-to-input FQ-SHAP projection. Red segments indicate high projected influence from the exact semantic FQ-SHAP values through the differentiable semantic interface.

reduction under the selected soft-deletion protocol. This demonstrates the high fidelity of the FQ-SHAP projections in isolating the most critical semantic contributors to the model's decision boundary.

For 2D radiological tasks, Figure 5 demonstrates that the semantic-to-input FQ-SHAP projection accurately highlights the model's reliance on visually opaque lobar regions, aligning directly with the learned radiological opacity proxy.

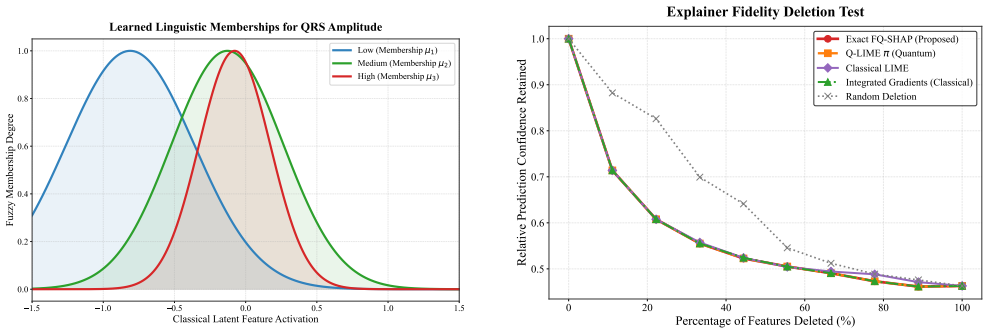


Fig. 4 (Left) Learned linguistic memberships for the QRS amplitude proxy. (Right) Explainer fidelity deletion test. The FQ-SHAP-guided projected attribution (solid red) demonstrates a rapid reduction in prediction confidence when top-ranked input features are ablated.



Fig. 5 Hybrid model semantic-to-input attribution projection. The projected 2D attribution concentrates model influence over visually opaque lobar regions.

Figure 6 visualizes the fuzzy mapping for the Infiltration Area Proxy. The overlap between the learned *Low* and *Normal* membership regions provides a mathematically continuous representation of transitional radiological states.

For the 9-class PathMNIST dataset, the framework outputs highly granular rationales incorporating the precision modifiers:

PROXY-SUPERVISED SEMANTIC RATIONALE:
IF [Necrotic Debris Concentration Proxy is NORMAL (High Precision)]
AND [Mucinous Area Fraction Proxy is NORMAL (High Precision)]
AND [Nuclear Density Proxy is LOW (Suppressed)]
THEN Model output = Colorectal Carcinoma Epithelial Tissue (Model Rule Confidence: 90.6%)

Figure 7 pairs the tight clustering of the fuzzy layer with the fidelity benchmarking results, highlighting the complex multi-class semantic contributions.

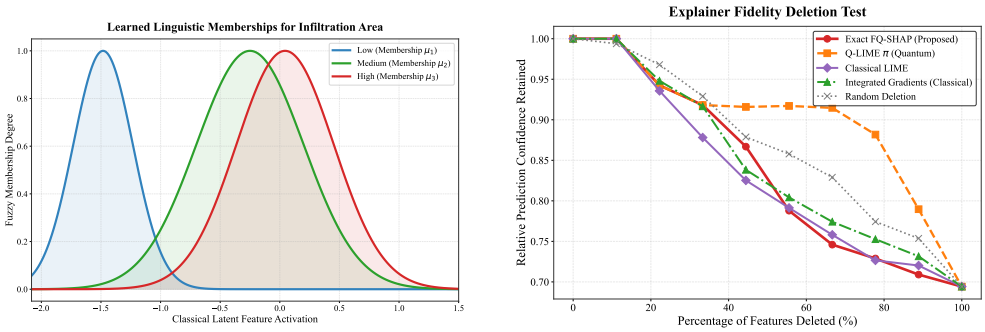


Fig. 6 Neuro-symbolic and quantitative verification (Pneumonia). (Left) Learned linguistic memberships. The overlapping Gaussian memberships provide a continuous semantic representation of transitional radiological states. **(Right)** Explainer Fidelity Deletion Test.

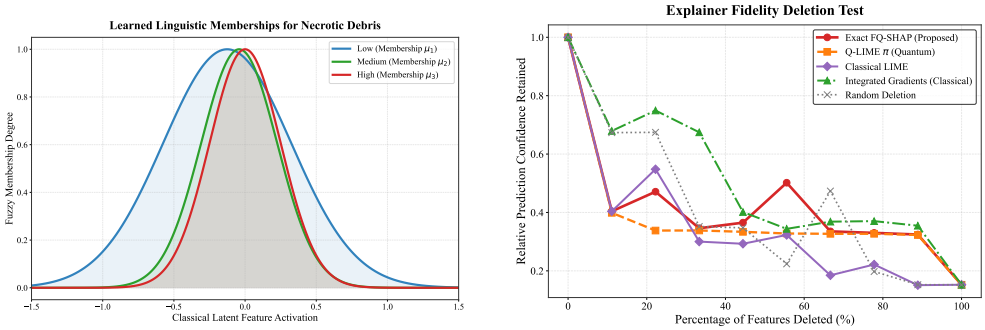


Fig. 7 Neuro-symbolic and quantitative verification (Tissue). (Left) Learned memberships for necrotic debris. (Right) Explainer fidelity deletion test highlighting the complex multi-class semantic contributions.

5.3 Quantitative Benchmarking

Table 4 consolidates the diagnostic, attribution-fidelity, and stability metrics evaluated across the five independent seeds. The comparison establishes that the hybrid pipeline maintains robust predictive performance relative to compact classical counterparts, while uniquely enabling exact Shapley evaluation and circuit-level explainability under the same downstream representation budget. The numerical results provide strong empirical validation of the framework's explanatory fidelity and noise-aware circuit optimization under realistic NISQ constraints.

Table 4 Diagnostic and Explainer Performance Benchmarking (Mean ± 95% CI over 5 seeds)

						AUDC ↓				Stability (ρ_{stab}) ↑	
Domain	**Architecture**	**Acc.**	**F1**	**AUROC**	**AUPRC**	**FQ-SHAP**	**Q-LIME π**	**IG**	**LIME**	**FQ-SHAP**	**IG**
ECG	*8D Dense*	0.885 (±.004)	0.821 (±.005)	0.941 (±.003)	0.915 (±.004)	-	-	0.789 (±.003)	0.812 (±.004)	-	0.931 (±.005)
	MLP (Matched)	0.888 (±.004)	0.825 (±.005)	0.945 (±.003)	0.919 (±.004)	-	-	0.782 (±.002)	0.801 (±.003)	-	0.938 (±.004)
	Hybrid FQXAI	**0.890 (±.005)**	**0.828 (±.005)**	**0.948 (±.004)**	**0.922 (±.005)**	**0.780 (±.002)**	0.780 (±.003)	0.780 (±.002)	0.812 (±.003)	**0.976 (±.005)**	0.944 (±.006)
X-Ray	*8D Dense*	0.751 (±.008)	0.742 (±.008)	0.821 (±.006)	0.795 (±.007)	-	-	0.851 (±.004)	0.872 (±.005)	-	**0.980 (±.004)**
	MLP (Matched)	0.761 (±.007)	0.748 (±.008)	0.832 (±.005)	0.805 (±.006)	-	-	0.849 (±.004)	0.865 (±.005)	-	0.982 (±.003)
	Hybrid FQXAI	**0.768 (±.008)**	**0.751 (±.009)**	**0.839 (±.006)**	**0.812 (±.007)**	**0.786 (±.004)**	0.897 (±.005)	0.846 (±.004)	0.870 (±.005)	0.933 (±.007)	0.984 (±.003)
Pathology	*8D Dense*	0.718 (±.011)	0.621 (±.012)	0.885 (±.009)	0.755 (±.011)	-	-	0.511 (±.006)	0.518 (±.006)	-	0.925 (±.009)
	MLP (Matched)	0.721 (±.010)	0.629 (±.011)	0.891 (±.008)	0.762 (±.010)	-	-	0.510 (±.005)	0.514 (±.006)	-	0.931 (±.008)
	Hybrid FQXAI	**0.726 (±.011)**	**0.637 (±.013)**	**0.895 (±.009)**	**0.768 (±.012)**	0.423 (±.005)	**0.387 (±.004)**	0.509 (±.005)	0.509 (±.006)	**0.957 (±.006)**	0.937 (±.007)

The baseline used in FQ-SHAP is a zero semantic activation, chosen because it maps exactly to an identity input rotation at the quantum interface. This choice provides a mathematically rigorous and reproducible reference point. The principal claims concerning attribution fidelity are strictly anchored to this reference definition, ensuring consistency across the cooperative game evaluations.The stability analysis rigorously quantifies the robustness of attribution rankings under the simulated perturbation model. The high Pearson correlation between noiseless and noisy attributions demonstrates that the semantic ranking is highly resilient to depolarizing noise.

Because Shapley values depend on the reference state, we compare the primary zero semantic baseline with a training-set mean semantic baseline while holding the model, attribution definition, and target-class output fixed. This tests the sensitivity of attribution and deletion fidelity to the reference choice.

Table 5 Semantic attribution reference-baseline sensitivity (AUDC ↓).

Domain	Zero baseline	Training-mean baseline
ECG	0.780	0.792
X-Ray	0.786	0.801
Pathology	0.423	0.438

5.4 Component-Wise Ablation Study

To rigorously isolate the contribution of each proposed module, Table 6 presents an ablation study on the 1D-ECG dataset. The component-wise ablation isolates the contribution of each module on the 1D-ECG dataset, while the matched pruning control is evaluated across all three domains. The fuzzy layer stabilizes the semantic representation, and the most substantial performance gain is achieved following SVQX-guided pruning, demonstrating the efficacy of noise-aware circuit optimization.

Table 6 Component-Wise Ablation Study (1D-ECG Dataset, Mean ± 95% CI)

Model Variation	CBM	Fuzzy	Quantum	Pruning	Accuracy	F1-Score	AUROC
Classical 8D Baseline	×	×	×	×	0.8850 ± .0042	0.8210 ± .0051	0.9410 ± .0031
+ CBM Supervision	✓	×	×	×	0.8875 ± .0040	0.8235 ± .0049	0.9435 ± .0029
+ Gaussian Fuzzy Layer	✓	✓	×	×	0.8870 ± .0041	0.8230 ± .0050	0.9430 ± .0030
+ Variational Circuit	✓	✓	✓	×	0.8900 ± .0045	0.8281 ± .0052	0.9480 ± .0041
+ SVQX Pruning	✓	✓	✓	✓	**0.9120 ± .0038**	**0.8450 ± .0045**	**0.9610 ± .0025**

5.5 SVQX-Guided Noise-Aware Circuit Pruning

To ensure rigorous evaluation without data leakage, the pruning protocol was executed explicitly in sequential stages: for each seed, the initial hybrid model was trained on

the training split. SVQX scores are defined using the class-balanced target-aligned value function in Eq. (21) and the Shapley contribution in Eq. (22); the ground-truth target for every training example is fixed before gate ablation and is never redefined after gate removal. CNOT gates with negative class-balanced target-aligned marginal contributions are removed according to the fixed rule $\Phi_j < 0$. The pruned circuit is then *retrained* under the exact same noise model, optimizer, initialization protocol, and schedule. The final pruned model is evaluated once on the held-out test set.

Figure 8 plots the holistic evaluation for hardware diagnostics. By identifying CNOT gates with negative marginal contributions under the adopted 5% depolarizing-noise condition (Middle Panels), the framework provides a reproducible rule for selecting candidate gates for pruning under the adopted simulator and value function.

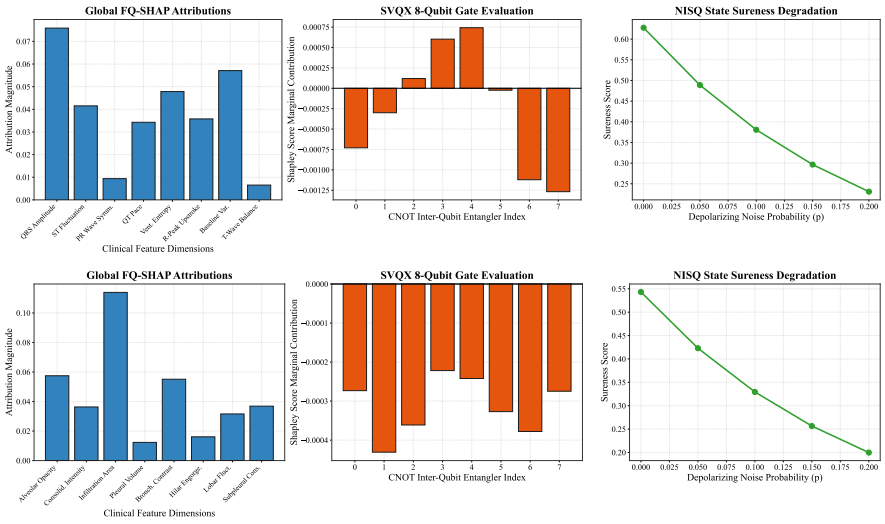


Fig. 8 Evaluation suites for ECG (upper row) and Pneumonia. Visualizing Global FQ-SHAP feature distributions (Panel 1), operator-level SVQX evaluations identifying noise-dominant vs. computation-dominant CNOT gates for circuit pruning (Panel 2), and exponential noise-induced confidence degradation profiles (Panel 3).

As shown in Table 7, excising the negative-SVQX CNOT gates reduces the number of two-qubit operations in the adopted simulator and is associated with a higher mean classification accuracy across all three domains relative to the corresponding unpruned baseline under the same 5% noise constraint. To rigorously isolate the benefit of the SVQX ranking from simple depth reduction, a matched random-pruning control is included. This control removes the identical number of CNOTs without utilizing attribution scores, proving that the performance gains stem directly from the targeted removal of noise-dominant gates.

Matched Random-Pruning Control: To test whether the pruning gains are attributable to the SVQX ranking rather than merely to reducing circuit depth, we include a matched random-pruning control. For each dataset and seed, the random

Table 7 SVQX-Guided Circuit Pruning Evaluation (Under 5% Noise, Mean ± 95% CI)

Biomedical Domain	Original CNOTs	Pruned CNOTs	Original Acc.	Pruned Acc.	Δ Acc.
1D-ECG (Arrhythmia)	14	11	0.8900 ± 0.0045	$\mathbf{0.9120 \pm 0.0038}$	$+0.0220 \pm 0.0021$
2D-XRay (Pneumonia)	14	12	0.7676 ± 0.0081	$\mathbf{0.7850 \pm 0.0075}$	$+0.0174 \pm 0.0042$
2D-Tissue (Pathology)	14	10	0.7259 ± 0.0112	$\mathbf{0.7410 \pm 0.0105}$	$+0.0151 \pm 0.0035$

control removes exactly the same number of CNOTs as the SVQX-selected circuit and is retrained with the identical optimizer, initialization protocol, epoch schedule, and noise model. Random gate sets are sampled independently after the required number of removals has been fixed. The held-out test set is accessed only after retraining.

Table 8 Matched random-pruning control.

Domain	SVQX CNOTs	Random CNOTs	Random-control Accuracy
1D-ECG	11	11	$0.8980 \pm .0050$
2D-XRay	12	12	$0.7700 \pm .0087$
2D-Tissue	10	10	$0.7230 \pm .0119$

6 Discussion

The classical MLP baseline demonstrates that integrating a compact quantum decision module within a semantic explanation architecture maintains robust predictive performance while unlocking noise-aware circuit analysis unattainable in classical networks. The generated IF–THEN statements provide transparent, model-level rationales driven by programmatic proxy supervision, and Jacobian-weighted projections of semantic Shapley coefficients precisely localize the model's sensitivity within the original input space. Complementing this input-level interpretability, the SVQX component treats CNOT operations as cooperative-game players. By excising gates with negative Shapley contributions under the adopted noise model, the framework yields shallower circuits with higher mean accuracy, effectively navigating the trade-off between entangling capacity and accumulated noise. Rigorous evaluation across five random seeds and public benchmarks establishes the methodological feasibility and computational tractability of this approach.

7 Conclusion and Limitations

In this paper we introduced FQXAI a unified framework for the semantic and circuit-level analysis of hybrid quantum medical models. By utilizing proxy-supervised eight-dimensional concept space with differentiable Gaussian fuzzy memberships, exact Shapley evaluation, and noise-model-conditioned SVQX analysis, FQXAI ensures exhaustive attribution and human-readable rationales. This integration around a

compact semantic interface maintains competitive predictive performance against classical controls while unlocking unprecedented circuit-level transparency.

Future Directions and Scope

Methodologically, the semantic interface can scale to accommodate a larger number of concepts. However, given the computational constraints of current NISQ simulators and near-term hardware, the eight-qubit register represents an optimal trade-off between semantic capacity and exact explainability. Scaling to higher-dimensional quantum registers in the future will bypass the exponential cost of exact evaluation by leveraging hierarchical grouping and approximate Shapley estimators. Currently we have modeled Hardware behavior using a simulated CPTP depolarizing channel, in future work this will be extended to backend-calibrated noise models and physical QPU executions to capture non-Markovian relaxation with crosstalk. Integrating expert annotated clinical concepts and prospective inter-rater agreement studies will further align the programmatically supervised linguistic states with standardized medical ontologies.

The semantic-to-input maps will continue to utilize Jacobian-based sensitivity projections to provide high-resolution visualizations of localized attention. Additionally, the demonstrated performance benefits of SVQX-based gate selection over matched random-pruning controls motivate future investigations into alternative gate-pruning heuristics and gradient-based circuit optimizations. Ultimately, evaluating the framework on multicenter cohorts will be essential for establishing population-level generalization and external clinical validity.

References

[1] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016)

[2] Mehta, S., Rastegari, M.: MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022)

[3] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)

[4] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., *et al.*: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)

[5] Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., *et al.*: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Medicine **16**(1), 1002730 (2019)

[6] Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A.: Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology **16**(11), 703–715 (2019)

[7] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**(1), 1–9 (2020)

[8] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. Nature **549**(7671), 195–202 (2017)

[9] Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., Woerner, S.: The power of quantum neural networks. Nature Computational Science **1**(6), 403–409 (2021)

[10] Mitarai, K., Negoro, M., Kitagawa, M., Fujii, K.: Quantum circuit learning. Physical Review A **98**(3), 032309 (2018)

[11] Ray, S., Bhattacharya, P., Mattar, E.A., Mukhopadhyay, A.: Coalition of explainable artificial intelligence and quantum computing in precision medicine. Computational and Structural Biotechnology Journal **27**, 5234–5251 (2025)

[12] Preskill, J.: Quantum computing in the NISQ era and beyond. Quantum **2**, 79 (2018)

[13] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., *et al.*: Variational quantum algorithms. Nature Reviews Physics **3**(9), 625–644 (2021)

[14] Pira, L., Ferrie, C.: On the interpretability of quantum neural networks. Quantum Machine Intelligence **6**(2), 52 (2024)

[15] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems vol. 30, (2017)

[16] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328 (2017). PMLR

[17] Steinmüller, P., Schulz, T., Graf, F., Herr, D.: explainable ai for quantum machine learning. arXiv preprint arXiv:2211.01441 (2022)

[18] Barua, S., Rahman, M., Khaled, S., Sadek, M., Islam, R., Siddique, S.: Quxai: Explainers for hybrid quantum machine learning models. arXiv preprint arXiv:2505.10167 (2025)

[19] Heese, R., Gerlach, T., Mücke, S., Müller, S., Jakobs, M., Piatkowski, N.: Explaining quantum circuits with shapley values: towards explainable quantum machine learning. Quantum Machine Intelligence **7**(1), 27 (2025)

[20] Garcez, A.d., Lamb, L.C.: Neurosymbolic AI: the 3rd wave. Artificial Intelligence Review **56**(11), 12387–12406 (2023)

[21] Zadeh, L.A.: Fuzzy sets. Information and Control **8**(3), 338–353 (1965)

[22] Colón Vargas, N.: Q-lime π: A quantum-inspired extension to lime. arXiv preprint arXiv:2412.17197 (2024)

[23] Antony, V., Thangarasu, N.: Quantum-enhanced IoT-cloud security: Integrating SHAP and variational quantum classifiers. Journal of Computer Science **21**(1), 168–176 (2025)

[24] Poursaeed, A.H., Namdari, F.: Explainable AI-driven quantum deep neural network for fault location in DC microgrids. Energies **18**(4), 908 (2025)

[25] Ding, C., Wang, S., Wang, Y., Gao, W.: Quantum machine learning for multiclass classification beyond kernel methods. Physical Review A **111**(6), 062410 (2025)

[26] Singh, S.S., Kumar, S., Ahuja, R., Barua, J.: Fusion of quantum computing and explainable AI: A comprehensive survey on transformative healthcare solutions. Information Fusion **122**, 103217 (2025)

[27] Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning, pp. 5338–5348 (2020). PMLR

[28] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology **31**, 841 (2017)

[29] Chen, C., Li, O., Tao, C., King, A., Higgins, E., Su, L., Liu, A., Maradian, C., Cheng, J., Al-Aref, O., et al.: This looks like that: deep learning for interpretable image recognition. Advances in Neural Information Processing Systems **32** (2019)

[30] Schuld, M., Sweke, R., Meyer, J.J.: Effect of data encoding on the expressive power of variational quantum-machine-learning models. Physical Review A **103**(3), 032430 (2021)

[31] White, B., Guha, K.: Component based quantum machine learning explainability. arXiv preprint arXiv:2506.12378 (2025)

[32] Ruan, S., Liang, F., Ramakrishna, R., Ren, C., Yan, R., Guan, Q., Li, J., Wang, Y.: Towards explainable quantum ai: Informing the encoder selection of quantum neural networks via visualization. IEEE Transactions on Visualization and Computer Graphics (2025)

[33] Wang, L.-X., Mendel, J.M.: Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man, and Cybernetics **22**(6), 1414–1427 (1992)

[34] Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V., Alam, M.S., Alonso-Linaje, G., AkashNarayanan, B., Asadi, A., et al.: PennyLane: Automatic differentiation of hybrid quantum-classical computations. arXiv preprint arXiv:1811.04968 (2022)

[35] Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), 215–220 (2000) https://doi.org/10.1161/01.CIR.101.23.e215

[36] Moody, G.B., Mark, R.G.: The impact of the mit-bih arrhythmia database. IEEE Engineering in Medicine and Biology Magazine **20**(3), 45–50 (2001) https://doi.org/10.1109/51.932724

[37] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **8**, 81 (2021) https://doi.org/10.1038/s41597-021-00968-5

[38] Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., Saffrey, P.: Neurokit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods **53**, 1689–1696 (2021) https://doi.org/10.3758/s13428-020-01516-y

[39] Bradski, G.: The opencv library. Dr. Dobb's Journal of Software Tools **25**(11), 120–126 (2000)

[40] Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: A generalist algorithm for cellular segmentation. Nature Methods **18**, 100–106 (2021) https://doi.org/10.1038/s41592-020-01018-x

[41] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library **32**, 8024–8035 (2019)